Chapter 12

Section 12.1

Check Your Understanding, page 752:

State: We want to estimate the true slope β of the population regression line relating fat gain to change in NEA with 95% confidence. *Plan*: We will use a *t* interval for the slope if the conditions are met. Linear: There is no leftover pattern in the residual plot, indicating that a linear model is appropriate. Independent: Knowing the fat gain for one subject should not help us predict the fat gain for another subject. Also, the sample size ($n = 16$) is less than 10% of all healthy young adults. Normal: The histogram of the residuals shows no strong skewness or outliers. Equal SD: Other than one point with a large positive residual, the residual plot shows roughly equal scatter for all *x* values. Random: These data come from a random sample. The conditions are met. *Do*: With df = 16 – 2 = 14, the confidence interval is: –0.0034415 ± 2.145(0.0007414) = –0.0034415 ± 0.00159 = (–0.005032, –0.001852). *Conclude*: We are 95% confident that the interval from –0.005032 to –0.001852 captures the slope of the population regression line relating fat gain to change in NEA.

Check Your Understanding, page 757:

State: We want to perform a test of $H_0 : \beta = 0$ versus $H_a : \beta < 0$ where β is the slope of the true regression line relating fat gain to NEA change. We will use $\alpha = 0.05$. *Plan*: We assume the conditions for inference are met and use a *t* test for the slope β. *Do*: According to the output, the test statistic is $t = -4.64$. Because the computer output includes a two-sided *P*-value, we must divide it by 2 to obtain the correct one-sided *P*-value: *P*-value $\approx 0.000/2 \approx 0$. *Conclude*: Because the *P*-value of approximately 0 is less than $\alpha = 0.05$, we reject H_0. There is convincing evidence that the slope of the true regression line relating fat gain to NEA change is negative.

Exercises, page 759:

12.1 The Equal SD condition is not met because the standard deviation of the residuals clearly increases as the laboratory measurement (*x*) increases.

12.2 The Linear condition is not met. There is clear curvature in the residual plot, which suggests that the relationship between mean SAT score and percent taking is not linear.

12.3 Linear: There is no leftover pattern in the residual plot, indicating that a linear model is appropriate. Independent: Knowing the BAC for one subject should not help us predict the BAC for another subject. Normal: The histogram of the residuals shows no strong skewness or outliers. Equal SD: The residual plot shows roughly equal scatter for all *x* values. Random: These data come from a randomized experiment. The conditions are met.

12.4 Linear: There is no leftover pattern in the residual plot, indicating that a linear model is appropriate. Independent: Knowing the proportion of perch killed in one pen should not help us predict the proportion of perch killed in another pen. Normal: The histogram of the residuals shows no strong skewness or outliers. Equal SD: The residual plot shows roughly equal scatter for all *x* values. Random: These data came from a randomized experiment. The conditions are met.

12.5 α is the true *y* intercept, which measures the true mean BAC level if no beers had been drunk. From the computer output, our estimate of α is $a = -0.012701$. β is the true slope, which measures how much the true mean BAC changes with the drinking of one additional beer. From the computer output, our estimate of β is $b = 0.018$. Finally, σ is the true standard deviation of the residuals, which measures how much the observed values of BAC typically vary from the population regression line. From the computer output, our estimate of σ is $s = 0.0204$.

12.6 α is the true *y* intercept, which measures the true average proportion of perch killed if there were 0 fish in the tank to begin with. In this case, the interpretation of the *y* intercept has no statistical meaning—if there were 0 perch in the pen, then there would be no perch killed. From the computer output, our estimate of α is $a = 0.12049$. β is the true slope, which measures how much the true average proportion of fish killed changes for each additional perch in the pen. From the computer output, our estimate of β is $b = 0.008569$. Finally, σ is the true standard deviation of the residuals, which measures how much the proportion killed values typically vary from the population regression line. From the computer output, our estimate of σ is $s = 0.1886$.

12.7 (a) The standard error of the slope is $SE_b = 0.0024$. If we repeated the experiment many times, the slope of the sample regression line would typically vary by about 0.0024 from the slope of the true regression line for predicting BAC from the number of beers consumed.
(b) With df = 16 – 2 = 14 and 99% confidence, $t^* = 2.977$. This leads to the confidence interval of $0.018 \pm 2.977(0.0024) = 0.018 \pm 0.007 = (0.011, 0.025)$.
(c) We are 99% confident that the interval from 0.011 to 0.025 captures the slope of the true regression line for predicting BAC from the number of beers consumed.
(d) If we repeated the experiment many times and computed a confidence interval for the slope each time, about 99% of the resulting intervals would contain the slope of the true regression line for predicting BAC from the number of beers consumed.

12.8 (a) The standard error of the slope is $SE_b = 0.002456$. If we repeated the experiment many times, the slope of the sample regression line would typically vary by about 0.002456 from the slope of the true regression line for predicting the proportion of perch killed from the number of perch in the pen.
(b) With df = 16 – 2 = 14 and 90% confidence, $t^* = 1.761$. This leads to the confidence interval of $0.00857 \pm 1.761(0.002456) = 0.00857 \pm 0.00433 = (0.00424, 0.0129)$.
(c) We are 90% confident that the interval from 0.00424 to 0.0129 captures the slope of the true regression line for predicting the proportion of perch killed from the number of perch in the pen.
(d) If we repeated the experiment many times and computed a confidence interval for the slope each time, about 90% of the resulting intervals would contain the slope of the true regression line for predicting the proportion of perch killed from the number of perch in the pen.

12.9 *State*: We want to construct a 99% confidence interval for β = the slope of the population regression line relating number of clusters of beetle larvae to number of stumps. *Plan*: We assume the conditions for inference are met and use a *t* interval for the slope to estimate β. *Do*: With df = 23 – 2 = 21, the confidence interval is $11.894 \pm 2.831(1.136) = 11.894 \pm 3.216 = (8.678, 15.11)$. *Conclude*: We are 99% confident that the interval from 8.678 to 15.11 captures the slope of the population regression line relating number of clusters of beetle larvae to number of stumps.

12.10 *State*: We want to construct a 90% confidence interval for β = the slope of the population regression line relating heights to arm spans for students in this high school. *Plan*: We assume the conditions for inference are met and use a *t* interval for the slope to estimate β. *Do*: With df = 18 – 2 = 16, the confidence interval is $0.8404 \pm 1.746(0.0809) = 0.8404 \pm 0.1413 = (0.6991, 0.9817)$. *Conclude*: We are 90% confident that the interval from 0.6991 to 0.9817 captures the slope of the population regression line relating heights to arm spans for students in this high school.

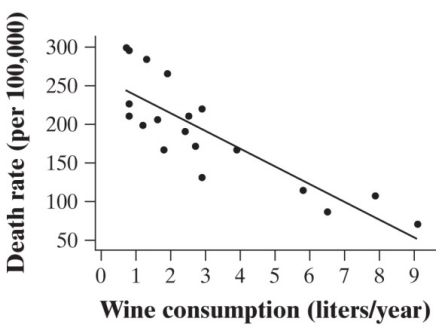
12.11 (a) $\hat{y} = -1.286 + 11.894(5) = 58.184$ clusters.
(b) From the computer output, our estimate of σ is s = 6.419. So, we would expect our prediction to be off from the actual number of clusters by about 6.419 clusters.

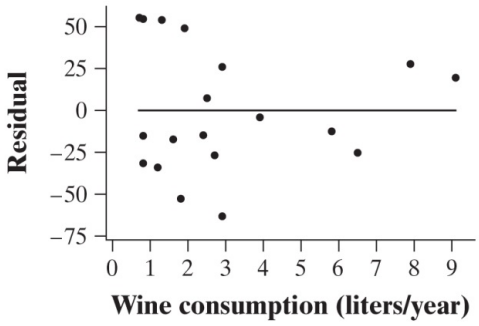
12.12 (a) $\hat{y} = 11.547 + 0.84042(76) = 75.4189$ inches.
(b) From the computer output, our estimate of σ is s = 1.613. So, we would expect our prediction to be off from the actual height by about 1.613 inches.

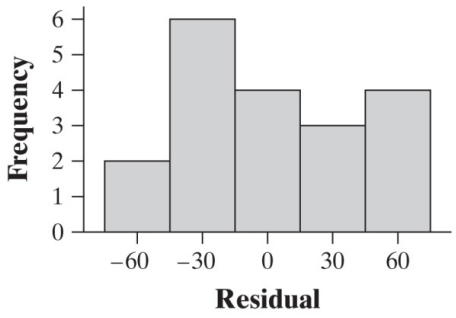
12.13 (a) The equation for the line is $\hat{y} = 166.483 - 1.0987x$ where $\hat{y}$ is the predicted corn yield and x is the number of weeds per meter. The slope says that for each additional weed per meter, the predicted corn yield will decrease by about 1.0987 bushels/acre. The y intercept says that if there are no weeds per meter, we would predict a corn yield of 166.483 bushels/acre.
(b) When using weeds per meter to predict corn yield, the actual yield will typically vary from the predicted yield by about 7.98 bushels/acre.
(c) *State*: We want to perform a test of $H_0 : \beta = 0$ versus $H_a : \beta < 0$ where β is the slope of the true regression line relating corn yield to weeds per meter. We will use $\alpha = 0.05$. *Plan*: We assume the conditions for inference are met and use a *t* test for the slope β. *Do*: According to the output, the test statistic is t = –1.92. Because the computer output includes a two-sided *P*-value, we must divide it by 2 to obtain the correct one-sided *P*-value. *P*-value = 0.075/2 = 0.0375. *Conclude*: Because the *P*-value of 0.0375 is less than α = 0.05, we reject H_0. There is convincing evidence that the slope of the true regression line relating corn yield to weeds per meter is negative. In other words, we have convincing evidence that having more weeds reduces corn yield.

12.14 (a) The equation for the line is $\hat{y} = 560.65 - 3.0771x$ where $\hat{y}$ is the predicted calories consumed and x is the time spent at the table. The slope says that for each additional minute at the table, the predicted number of calories will decrease by about 3.0771 calories. The y intercept says that if there no time spent at the table, the predicted number of calories consumed would be about 560.65. Because it is impossible for a child to eat if he or she is at the table for 0 minutes, interpreting the y intercept doesn't make sense in this case.
(b) When using time at the table to predict number of calories, the actual number of calories will typically vary from the predicted calories by about 23.3980 calories.
(c) *State*: We want to perform a test of $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$ where β is the slope of the population regression line relating calorie consumption to time at the table in the population of toddlers. We will use $\alpha = 0.01$. *Plan*: We assume the conditions for inference are met and use a *t* test for the slope β. *Do*: According to the output, the test statistic is –3.62 and the two-sided *P*-value is 0.002. *Conclude*: Because the *P*-value of 0.002 is less than α = 0.01 we reject H_0. There is convincing evidence of a linear relationship between time at the table and calorie consumption in the population of toddlers.

12.15 (a) *State*: We want to perform a test of $H_0 : \beta = 0$ versus $H_a : \beta < 0$ where β is the slope of the population regression line relating heart disease death rate to wine consumption in the population of countries. We will use $\alpha = 0.05$. *Plan*: If the conditions are met, we will do a *t* test for the slope β. Linear: There is no leftover pattern in the residual plot (shown below), indicating that a linear model is appropriate. Independent: Knowing the heart disease death rate for one country should not help us predict the heart disease death rate for another country. Also, the sample size (n = 19) is less than 10% of all countries. Normal: The histogram of residuals (shown below) shows no strong skewness or outliers. Equal SD: The residual plot shows that the standard deviation of the death rates might be a little smaller for large values of wine consumption x, but it is hard to tell with so few data values. Random: The data come from a random sample. Because the Equal SD condition is questionable, we will proceed with caution.

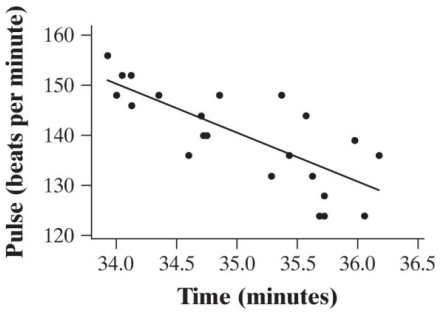


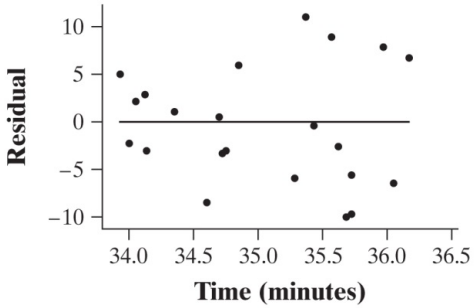


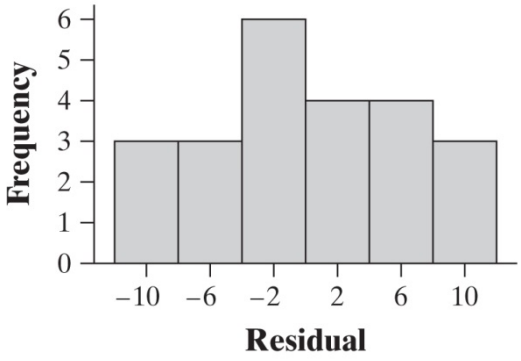


Do: Using technology, $t = -6.46$, df = 19 – 2 = 17, and *P*-value ≈ 0. *Conclude*: Because the *P*-value of approximately 0 is less than $\alpha = 0.05$, we reject H_0. There is convincing evidence of a negative linear relationship between wine consumption and heart disease death rate in the population of countries.

12.16 (a) *State*: We want to perform a test of $H_0 : \beta = 0$ versus $H_a : \beta < 0$ where β is the slope of the population regression line relating pulse rate to swim time for Professor Moore. We will use $\alpha = 0.05$. *Plan*: If the conditions are met, we will do a *t* test for the slope β. Linear: There is no leftover pattern in the residual plot (shown below), indicating that a linear model is appropriate. Independent: Knowing the pulse rate for one swim should not help us predict the pulse rate for another swim. Also, the sample size (n = 23) is less than 10% of all days when Professor Moore swims 2000 yards. Normal: The histogram of residuals (shown below) shows no strong skewness or outliers. Equal SD: The residual plot shows that the standard deviation of the pulse rates might be a little smaller for small values of x, but it is hard to tell with so few data values. Random: The data come from a random sample. Because the Equal SD condition is questionable, we will proceed with caution.







Do: Using technology, $t = -5.13$, df = 23 – 2 = 21, and *P*-value ≈ 0. *Conclude*: Because the *P*-value of approximately 0 is less than $\alpha = 0.05$, we reject H_0. There is convincing evidence of a negative linear relationship between swim time and pulse rate in the population of days when Professor Moore swims 2000 yards.

12.17 (a) With df = 21 – 2 = 19, the confidence interval is
$11{,}630.6 \pm 2.093(1249) = 11{,}630.6 \pm 2614.2 = (9016.4, 14{,}244.8)$.

(b) Because the automotive group claims that people drive 15,000 miles per year, that says that for every increase of 1 year, the mileage would increase by 15,000 miles. In other words, the line relating miles to years would have a slope of 15,000 miles/year.

(c) $t = \dfrac{11{,}630.6 - 15{,}000}{1249} = -2.70$. With df = 21 – 2 = 19, the *P*-value is between 2(0.005) = 0.01 and 2(0.01) = 0.02. *Using technology*: *P*-value = 0.0142. Because the *P*-value of 0.0142 is less than α = 0.05, we reject H_0. We have convincing evidence that the slope of the population regression line relating miles to years is not equal to 15,000.
(d) Yes. Because the interval in part (a) does not include the value 15,000, we reject H_0. The interval also provides convincing evidence that the slope of the population regression line relating miles to years is not equal to 15,000.

12.18 (a) With df = 16 – 2 = 14, the confidence interval is
$0.79021 \pm 2.977(0.07104) = 0.79021 \pm 0.211486 = (0.5787, 1.0017)$.
(b) Both variables are measuring tire wear in the same units. If one method of measuring wear gives an increase in wear of 1 unit, we expect that the other way of measuring wear would also give an increase in wear of 1 unit. This translates into a slope of 1/1 = 1.
(c) $t = \dfrac{0.79021 - 1}{0.07104} = -2.95$. With df = 16 – 2 = 14, the *P*-value is between 2(0.005) = 0.01 and 2(0.01) = 0.02. *Using technology*: *P*-value = 0.0105. Because the *P*-value of 0.0105 is greater than α = 0.01, we fail to reject H_0. There is not convincing evidence that the slope of the population line relating wear measurements using the groove method to wear measurements using the weight method is different from 1.
(d) Yes. Because the interval in part (a) includes the value 1, we fail to reject H_0. The interval also does not provide convincing evidence that the slope of the population line relating wear measurements using the groove method to wear measurements using the weight method is different from 1. However, the values might not be the same even if the slope is 1. For example, if the groove method was always 5000 miles less than the weight method, the slope would still be 1 (groove = –5000 + weight).

12.19 c

12.20 c

12.21 a

12.22 e

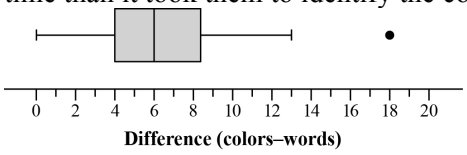
12.23 b

12.24 a

12.25 (a) This was an experiment because the two treatments (say the color of the printed word and read the word) were deliberately assigned to the students.
(b) He used a randomized block design where each student was a block. In other words, he used a matched pairs design because there were only two treatments per block. He did this to help account for the different abilities of students to read the words or to say the color they were printed in.

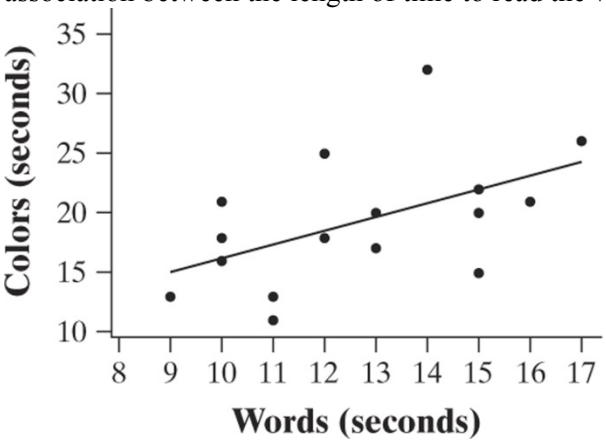
(c) The random assignment was used to help average out the effects of the order in which people did the two treatments. For example, if every subject said the color of the printed word first and were frustrated by this task, the times for the second treatment (reading the word) might be worse. Then we wouldn't know the reason the times were longer for the second treatment—because of frustration or because the second method actually takes longer.

12.26 First, calculate the difference in times for each student (we used Colors – Words). A boxplot of the differences is shown below. The median difference is 6 and the minimum is 0. For all students, the time it took to identify the color was the same or longer than the time needed to read the words. Typically, the students were able to read the words in about 6 seconds less time than it took them to identify the color.



12.27 It is not safe to use a paired *t* procedure because there are a small number of differences ($n_d = 16 < 30$) and there is an outlier.

12.28 (a) The scatterplot is below. There appears to be a moderately strong, positive linear association between the length of time to read the word and length of time to identify the color.



(b) The regression equation is $\hat{y} = 4.887 + 1.1321x$ where $\hat{y}$ is the predicted time to identify the color and x is the amount of time to read the words.

(c) The predicted value for the student who completed the word task in 9 seconds is $\hat{y} = 4.887 + 1.1321(9) = 15.076$ seconds, so the residual is $y - \hat{y} = 13 - 15.076 = -2.076$ seconds. This student took 2.076 seconds less than expected to identify the colors, based on his or her time to read the words.

(d) If the true slope of the regression line relating time to identify the colors and time to read the words is 0 and this experiment were repeated many times, there is a 0.0215 probability of getting an observed slope of 1.1321 or larger by chance alone.

12.29 (a) (i) The probability that the person is a snowmobile owner is $\frac{295}{1526} = 0.1933$.

(ii) The probability that the person belongs to an environmental organization or owns a snowmobile is $\frac{295 + 77 + 212}{1526} = 0.3827$.

(iii) The probability that the person has never used a snowmobile given that they belong to an environmental organization is $\frac{212}{305} = 0.6951$.

(b) No. The probability that a person is a snowmobile owner (295/1526 = 0.1933) is different than the probability that the person is a snowmobile owner given that he or she belong to an environmental organization (16/305 = 0.0525). If a person belongs to an environmental organization, they are more likely to have never used a snowmobile.

(c) (i) $P(\text{both are owners}) = \left(\frac{295}{1526}\right)\left(\frac{294}{1525}\right) = 0.0373$. There is a 0.0373 probability that both are owners.

(ii) P(at least one belongs to an environmental organization) = 1 – P(neither belong) = $1 - \left(\frac{1221}{1526}\right)\left(\frac{1220}{1525}\right) = 0.3599$. There is a 0.3599 probability that at least one of the two belongs to an environmental organization.

12.30 *State*: We want to perform a test of H_0: There is no association between environmental club membership and snowmobile use for the population of visitors to Yellowstone National Park versus H_a: There is an association between environmental club membership and snowmobile use for the population of visitors to Yellowstone National Park at the $\alpha = 0.05$ level. *Plan*: We should use a chi-square test for independence if the conditions are met. Random: The data come from a random sample. 10%: The sample size (n = 1526) is less than 10% of the population of visitors to Yellowstone National Park. Large Counts: The expected counts are all at least 5 (see table below).

Snowmobile use	No	Yes
Never used	525.69	131.31
Snowmobile renter	459.28	114.72
Snowmobile owner	236.04	58.96

Do: The test statistic is $\chi^2 = \frac{(445 - 525.69)^2}{525.69} + \ldots + \frac{(16 - 58.96)^2}{58.96} = 116.588$. With df = (3 – 1)(2 – 1) = 2, the P-value is approximately 0. *Conclude*: Becasue the P-value of approximately 0 is less than $\alpha = 0.05$, we reject H_0. There is convincing evidence that there is an association between environmental club membership and snowmobile use in the population of visitors to Yellowstone National Park.